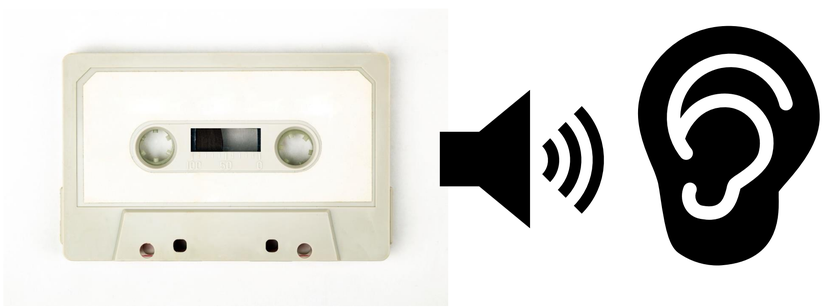
LINNAEUS UNIVERSITY, FACULTY OF ARTS AND HUMANITIES / DEPARTMENT OF CULTURAL SCIENCES

Automatic subject indexing of oral history interviews with Whisper and Claude

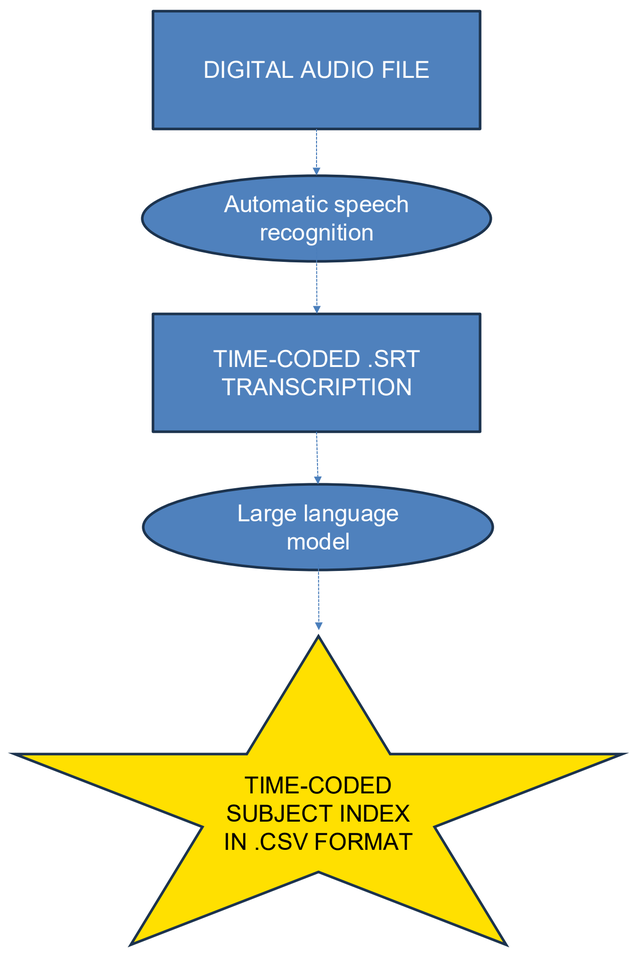
Oral history interviews are invaluable pieces of cultural memory, often freely available in online databases. But if you are looking for interviews discussing a particular topic, how do you find them?



OLD WAY

LISTEN FROM START TO FINISH

NEW WAY



PILOT STUDY:
14 oral history interviews in Swedish/South Sámi

BACKGROUND

The Sámi archival material in the DAUM collections contains numerous recordings, in Swedish and Sámi languages, that present Sámi counter-narratives to the dominant Swedish history (Cocq, 2010). Making collections like these accessible and searchable for a general audience requires a tremendous indexing effort, presenting large financial obstacles (Lambert, 2023).

PURPOSE AND GOAL

This pilot study investigates the feasibility and accuracy of automatically transcribing and indexing oral history interviews from the DAUM collections using OpenAI's Whisper and Claude 3.5 Sonnet from Anthropic.

METHOD

Fourteen recordings in Swedish mixed with South Sámi from the DAUM collections were transcribed using Whisper. The resulting time-coded srt-files were passed to Claude's API using the following prompt:

”Describe which subjects are covered in which parts of the following interview. Specify exact times.
Format the result as a CSV file with semicolon as separator and with two columns for time, as in this example:
From;To;Subject
00:00:00,000;00:00:11,000;A terrible storm”

The resulting time-coded subject indexes were manually checked for accuracy and compared with the content notes provided by DAUM.

RESULTS

The quality of the transcription depends on several factors, such as the sound quality, accents of the speakers and amount of language mixing. Audio segments in South Sámi were naturally misinterpreted, confusing the subsequent subject indexing. Overall, however, this method produced highly useful indexes giving functional indications on what to find in the different parts of the recordings.

From	To	Subject
00:00:00,000	00:00:29,000	Discussion about the person's origins from Marskjell and Jämtland
00:00:29,000	00:01:03,000	Dialogue about language differences and children's language learning
00:01:03,000	00:02:27,000	Discussion about the school system - mission schools and nomad schools
00:02:27,000	00:03:14,000	Details about the school year and terms
00:03:14,000	00:04:20,000	Teaching and subjects in school
00:04:20,000	00:05:27,000	Christian education and textbooks
00:05:27,000	00:06:12,000	Life after school and personal history
00:06:12,000	00:08:15,000	Work as a reindeer herder and personal background
00:08:15,000	00:09:21,000	Experiences from reindeer herding and working conditions
00:09:21,000	00:13:56,000	Discussion about language and education in Sami
00:13:56,000	00:19:27,000	Reindeer herding, seasons and migrations
00:19:27,000	00:22:28,000	Trade and food supply
00:22:28,000	00:23:41,000	Conversation about Sami language and greeting phrases

REFERENCES

1 Cocq, C. (2010). Forskningshistoriskt perspektiv på insamlingen av samiskt arkivmaterial. In Reinhammar, M. (ed.). *Svenska landsmål och svenskt folkliv 2010*. 133. H. 336.
2 Lambert, D. (2023). Oral History Indexing. *The Oral History Review*, 50(2), 169–192. https://doi.org/10.1080/00940798.2023.2235000

Linnæus University

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation

Faculty of Arts and Humanites/
Department of Cultural Sciences
Linnaeus University, Växjö, Sweden
Johannes Widegren
johannes.widegren@lnu.se